LABORATOR #10

EX#1 (Covarianță și corelație)

Covarianța și corelația a două variabile aleatoare X și Y sunt definite ca:

$$\mathrm{Cov}(X,Y) \coloneqq \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right];$$
$$\mathrm{Corr}(X,Y) \coloneqq \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)},$$

unde reamintim definiția deviației standard: $\sigma(X) \coloneqq \sqrt{\mathrm{Var}(X)} \coloneqq \sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]}$.
Observații:

- Dacă X și Y sunt independente, atunci covarianța și corelația lor sunt egale cu zero.
- Corelația este o mărime adimensională și ia valori în intervalul $[-1,1]$.

Aproximarea numerică a covarianței (Covarianța Eșantionului):
Fie (X_1,Y_1), (X_2,Y_2), ..., (X_N,Y_N) perechi independente de variabile aleatoare cu aceeași distribuție, având medie și varianță finite. Mai precis, pentru orice $a,b\in\mathbb{R}$ și $i\in\overline{2,N}$,

$$\mathbb{P}(X_i\le a\cap Y_i\le b) = \mathbb{P}(X_1\le a\cap Y_1\le b),$$

iar dacă $i\ne j$ și $a,a',b,b'\in\mathbb{R}$, atunci avem:

$$\mathbb{P}((X_i\le a\cap Y_i\le b)\cap(X_j\le a'\cap Y_j\le b')) = \mathbb{P}(X_i\le a\cap Y_i\le b)\,\mathbb{P}(X_j\le a'\cap Y_j\le b').$$

Dacă notăm

$$\mathrm{Cov}_N \coloneqq \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X}_N)(Y_i-\bar{Y}_N),$$

atunci

$$\mathrm{Cov}_N \xrightarrow{\text{aproape sigur}} \mathrm{Cov}(X_1,Y_1)$$

și

$$\mathbb{E}[\mathrm{Cov}_N] = \mathrm{Cov}(X_1,Y_1).$$

#1 La aruncarea a două zaruri, considerăm următoarele variabile aleatoare:

- $X=$ valoarea primului zar
- $Y=$ suma celor două zaruri.

Aproximați empiric prin simulări repetate covarianța și corelația celor două variabile aleatoare.

#2 Fișierul `data_file.py` conține funcția `get_data(<key>)`. Apelând `get_data("geyser")`, obținem un dicționar ce conține la cheia `"eruptions"` un vector cu durata (în minute) a mai multor erupții ale unui gheizer din Parcul Național Yellowstone, SUA. La cheia `"waiting"` se află un vector de aceeași lungime ce conține timpul (în minute) până la următoarea erupție. Calculați covarianța și corelația dintre durata erupțiilor și timpul până la următoarea erupție.

#3 Aceeași funcție apelată de data aceasta `get_data("cars")` returnează un dicționar cu date despre mașini. La cheia `"tons"` se află un vector cu masa mașinilor, la cheia `"range"` se află un vector de aceeași lungime cu numărul de kilometri parcurși utilizând un litru de combustibil, iar la cheia `"hp"` se află un vector cu puterea mașinii exprimată în cai putere. Calculați covarianța și corelația fiecărei perechi de caracteristici ale mașinilor.

EX#2 (Regresia liniară – preziceri ale unor date necunoscute)

Regresia liniară simplă. Considerăm un set de date de forma $(x_i, y_i)_{i=1}^N$ și dorim să găsim o relația liniară care aproximează cel mai bine dependența caracteristicilor y_i de caracteristicile x_i. Prin urmare, vrem să găsim dreapta de ecuația $y = ax + b$ care aproximează *cel mai bine* setul de date, în sensul minimizării următoarei erori (numită eroarea în sensul *celor mai mici pătrate*):

$$\mathrm{Err}_{\mathrm{LS}}(a, b) := \sum_{i=1}^{N} (y_i - ax_i - b)^2.$$

Desfacem parantezele și obținem:

$$\mathrm{Err}_{\mathrm{LS}}(a, b) = a^2 \sum_{i=1}^{N} x_i^2 + Nb^2 + 2ab \sum_{i=1}^{N} x_i - 2a \sum_{i=1}^{N} x_i y_i - 2b \sum_{i=1}^{N} y_i.$$

În cazul nedegenerat când nu avem toate valorile x_i egale, funcția este strict convexă, deci minimul global coincide cu unicul punct critic. Prin derivare parțială, obținem:

$$\partial_a \mathrm{Err}_{\mathrm{LS}}(a, b) = 2a \sum_{i=1}^{N} x_i^2 + 2b \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} x_i y_i;$$
$$\partial_b \mathrm{Err}_{\mathrm{LS}}(a, b) = 2Nb + 2a \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} y_i.$$

Egalăm cele două derivate parțiale cu zero și obținem sistemul:

$$\begin{pmatrix} \sum\limits_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum\limits_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{N} x_i y_i \\ \sum\limits_{i=1}^{N} y_i \end{pmatrix}.$$

În cazul nedegenerat, matricea e inversabilă deci putem rezolva sistemul și determinăm valorile a și b.

Regresia liniară cu două variabile de intrare. Considerăm acum setul de date $(x_i, y_i, z_i)_{i=1}^N$ și vrem să aproximăm liniar dependența valorilor z_i de valorile x_i și y_i. Prin urmare, vrem să găsim coeficienții a, b, c astfel încât planul de ecuație $z = ax + by + c$ să aproximeze cel mai bine setul de date. Eroare în sensul celor mai mici pătrate devine:

$$\mathrm{Err}_{\mathrm{LS}}(a, b) := \sum_{i=1}^{N} (z_i - ax_i - by_i - c)^2.$$

Procedând analog, în cazul nedegenerat în care nu toate punctele (x_i, y_i) din setul de date sunt coliniare, obținem:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Z},$$

unde:

$$\mathbf{X} = \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{pmatrix} \quad \text{și} \quad \mathbf{Z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}.$$

#1 Afișați într-un sistem de coordonate xOy setul de date `"geyser"` de la Ex#1, împreună cu dreapta de regresie a timpului până la următoarea erupție în funcție de durata unei erupții.

Preziceți, folosind dreapta de regresie obținută, câte minute vom aștepta până la următoarea erupție, dacă erupția curentă a durat 6 minute.

#2 Aceeași cerință pentru setul de date `"cars"`, în care să afișați dreapta de regresie a distanței parcurse cu un litru de combustibil în funcție de puterea mașinii.

#3 Preziceți câți kilometri putem parcurge cu un rezervor plin cu 40 de litri de combustibil, pentru o mașină de 2 tone și 97 cai putere, folosind regresia liniară a setului de date `"cars"`.